Interaction between Colligation, Register and Surface Variability in German Learners and Natives

Hagen Hirschmann
hirschhx@rz.hu-berlin.de
Amir Zeldes
amir.zeldes@rz.hu-berlin.de
Anke Lüdeling
anke.luedeling@rz.hu-berlin.de

Humboldt-Universität zu Berlin

Research questions

- Do learners distinguish register?
- If so, how much?
- Similarly or differently to natives?
- What is particularly difficult for learners in the acquisition of registers?
Overview

- Studying learner language
- Operationalizing interlanguage differences quantitatively
- Case study: adverbs and adverb chains in L1 and L2 registers
Data for L2 Studies

- Intuition / introspection (learner or teacher)
- Questionnaires (Diehl et al. 1991)
- Corpus data:
  - Learner corpora (Pravec 2002; Tono 2003; Granger 2008) and comparable L1 corpora
  - Metadata – reference to L2 proficiency, learner’s L1...
  - Annotation – pos, lemmatization, possibly error annotation (Corder 1981; Granger 2008)
Working with raw learner data

- Frequencies of word forms, annotated categories, or colligations using both
  - Work on lexical density as an index of L2 competence (Halliday 1989; Laufer/Nation 1999)
  - Studies using underuse/overuse compared to native data in the framework of Contrastive Interlanguage Analysis (Selinker 1972; Ringbom 1998; Granger et al. 2002)
Underuse and Overuse

- Simplified model of target register competence
- Learner’s interlanguage distributions as opposed to L1 distributions
- Underuse and overuse defined as statistically significant deviations from L1 control frequencies
Underuse as an index of difficulty

- Phenomena that are underrepresented can either be:
  - Unknown to learners (e.g. probably the word *forthwith*)
  - Known but (more or less consciously) avoided (e.g. the *past perfect progressive*)
L1 Independence

- Some errors are strongly L1 dependent, i.e. transfer errors:
  
is beautiful! (Italian pro-drop transfer)

- We are interested in phenomena that apply to GFL learners independently of L1

- Use L1 metadata to rule out interference and other language dependent effects
Visualizing Underuse/Overuse

- Normalized frequencies can be collected:
  - lexical categories (lemmas)
  - grammatical categories (POS n-grams)
- Degree of deviation from native frequency is represented in progressively warmer or colder colors
Visualization of Lexical Data

<table>
<thead>
<tr>
<th>lemma</th>
<th>tot_norm</th>
<th>de</th>
<th>da</th>
<th>en</th>
<th>fr</th>
<th>pl</th>
<th>ru</th>
</tr>
</thead>
<tbody>
<tr>
<td>in</td>
<td>0.013188</td>
<td>0.012261</td>
<td>0.014041</td>
<td>0.014247</td>
<td>0.015272</td>
<td>0.012135</td>
<td>0.009534</td>
</tr>
<tr>
<td>es</td>
<td>0.010897</td>
<td>0.011945</td>
<td>0.010900</td>
<td>0.011379</td>
<td>0.013347</td>
<td>0.008163</td>
<td>0.012385</td>
</tr>
<tr>
<td>sie</td>
<td>0.010618</td>
<td>0.008193</td>
<td>0.010643</td>
<td>0.008835</td>
<td>0.010909</td>
<td>0.006067</td>
<td>0.005613</td>
</tr>
<tr>
<td>dass</td>
<td>0.009522</td>
<td>0.007404</td>
<td>0.012823</td>
<td>0.008789</td>
<td>0.009625</td>
<td>0.008880</td>
<td>0.009890</td>
</tr>
<tr>
<td>von</td>
<td>0.007982</td>
<td>0.007122</td>
<td>0.007309</td>
<td>0.006846</td>
<td>0.007315</td>
<td>0.010259</td>
<td>0.007930</td>
</tr>
<tr>
<td>auch</td>
<td>0.007028</td>
<td>0.008362</td>
<td>0.008527</td>
<td>0.005828</td>
<td>0.005775</td>
<td>0.005461</td>
<td>0.004455</td>
</tr>
<tr>
<td>für</td>
<td>0.006683</td>
<td>0.007201</td>
<td>0.006091</td>
<td>0.007216</td>
<td>0.006802</td>
<td>0.005736</td>
<td>0.004188</td>
</tr>
<tr>
<td>sind</td>
<td>0.006465</td>
<td>0.004271</td>
<td>0.008976</td>
<td>0.007308</td>
<td>0.006930</td>
<td>0.004964</td>
<td>0.005346</td>
</tr>
<tr>
<td>sich</td>
<td><strong>0.006309</strong></td>
<td><strong>0.011697</strong></td>
<td><strong>0.006283</strong></td>
<td><strong>0.006291</strong></td>
<td><strong>0.006930</strong></td>
<td><strong>0.007170</strong></td>
<td><strong>0.005435</strong></td>
</tr>
<tr>
<td>ich</td>
<td>0.006262</td>
<td>0.003877</td>
<td>0.013272</td>
<td>0.005366</td>
<td>0.003453</td>
<td>0.001434</td>
<td>0.001426</td>
</tr>
<tr>
<td>aber</td>
<td>0.006048</td>
<td>0.003347</td>
<td>0.007309</td>
<td>0.006245</td>
<td>0.007315</td>
<td>0.003365</td>
<td>0.003831</td>
</tr>
</tbody>
</table>

Reflexive *sich* ‘self’ is underused

Excel Under/Overuse Addin: [http://korpling.german.hu-berlin.de/~amir/uoaddin.htm](http://korpling.german.hu-berlin.de/~amir/uoaddin.htm)
Underuse of pos-chains in L2 data

<table>
<thead>
<tr>
<th>bigram</th>
<th>tot_norm</th>
<th>de</th>
<th>da</th>
<th>en</th>
<th>fr</th>
<th>pl</th>
<th>ru</th>
</tr>
</thead>
<tbody>
<tr>
<td>$.-PPER</td>
<td>0.042384</td>
<td>0.005297</td>
<td>0.009748</td>
<td>0.007963</td>
<td>0.006166</td>
<td>0.005801</td>
<td>0.007409</td>
</tr>
<tr>
<td>VVFIN-$,$</td>
<td>0.042131</td>
<td>0.006457</td>
<td>0.00776</td>
<td>0.006343</td>
<td>0.006937</td>
<td>0.006243</td>
<td>0.008391</td>
</tr>
<tr>
<td>PPOSAT-NN</td>
<td>0.041739</td>
<td>0.008058</td>
<td>0.007247</td>
<td>0.007269</td>
<td>0.007066</td>
<td>0.006298</td>
<td>0.005802</td>
</tr>
<tr>
<td>ADV-ADV</td>
<td>0.041604</td>
<td>0.012858</td>
<td>0.010518</td>
<td>0.006111</td>
<td>0.006166</td>
<td>0.003094</td>
<td>0.002856</td>
</tr>
<tr>
<td>ADV-APPR</td>
<td>0.039742</td>
<td>0.009117</td>
<td>0.008016</td>
<td>0.005324</td>
<td>0.007837</td>
<td>0.004807</td>
<td>0.004642</td>
</tr>
<tr>
<td>PDAT-NN</td>
<td>0.03956</td>
<td>0.005409</td>
<td>0.004233</td>
<td>0.005509</td>
<td>0.007837</td>
<td>0.007735</td>
<td>0.008837</td>
</tr>
<tr>
<td>ADV-ART</td>
<td>0.037125</td>
<td>0.007629</td>
<td>0.006349</td>
<td>0.006898</td>
<td>0.005653</td>
<td>0.006133</td>
<td>0.004463</td>
</tr>
</tbody>
</table>

Multiple adverb-chains are generally underused

used tagset: STTS; http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide
ADVs in registers and learner language

- ADV-underuse characteristic of advanced learner variety
- Biber 2009: adverb type and token frequencies relevant for measuring register differences
- Independent or interacting factors?
Registers in L2 data?

- "lack of register awareness" (Gilquin/Paquot 2007)
- this predicts:
  - a general underuse of ADVs and ADV chains
  - no significant ADV-differences between registers
- Production of L2 ADV-ADV-chains dependent on (syntactic) complexity
  (Zeldes, Hirschmann & Lüdeling 2008)
Study/approach

a) Comparing ADV-n-grams:
   - ADV
   - ADV-ADV
   - ADV-ADV-ADV

   in L1 and L2 data with different registers

b) Comparing different syntactic structures
    of consecutive ADVs
Corpora for this study

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>academic theses</td>
<td>1,804,993</td>
<td></td>
</tr>
<tr>
<td>law texts</td>
<td>5,896,940</td>
<td></td>
</tr>
<tr>
<td>Falko Essays L1</td>
<td>67,529</td>
<td>Falko Essays L2</td>
</tr>
<tr>
<td>Falko Summaries L1</td>
<td>21,211</td>
<td>Falko Summaries L2</td>
</tr>
<tr>
<td>parliament debates</td>
<td>36,723,139</td>
<td></td>
</tr>
</tbody>
</table>

For an introduction to the Falko corpus, Lüdeling et al. 2008
Study a):
ADV-n-gram comparison

Raw L1 data:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Falko Essays L1</td>
<td>881.8</td>
</tr>
<tr>
<td>parliament debates</td>
<td>514.8</td>
</tr>
<tr>
<td>Falko Summaries L1</td>
<td>450.7</td>
</tr>
<tr>
<td>academic theses</td>
<td>215.6</td>
</tr>
<tr>
<td>law texts</td>
<td>189.6</td>
</tr>
</tbody>
</table>

numbers normalized to 10,000 tokens
ADV-ADV-chains (L1)

- essays L1
- parliament deb.
- summ. L1
- acad. theses
- law texts
ADV-ADV-chains (incl. L2)

- essays L1
- essays L2
- parliament deb.
- summ. L1
- summ. L2
- acad. theses
- law texts
Study b):
Different types of ADV-ADV-bigrams

- Method:
  - Syntactically classify ADV-ADV-bigrams
  - Token frequencies for each class from a Treebank (Tiger)
  - Compare frequencies in L1 & L2 registers
Category [ADV-ADV]

- Least complex category
- Lexicalized pairs (*immer noch* – still) or left headed (*morgen früh* – tomorrow early) or right headed (*sehr bald* – very soon) AdvPs
- Temporal adverbials, local adverbials
Category [ADV][ADV+X]

- sentence adverb, (coincidentally) followed by phrase-internal adverb
- More complex category
- Many focus particles & intensifiers
Category [ADV][ADV]

- Independent consecutive sentence adverbs
- Most complex category
- Many modal particles

if this certainly so works out is the question
ADV-ADV-types

- complex;
+ lex.;
+ tempor./local Adv.

+/- complex;
+ focus p. & intensifiers

+ complex;
+ modal p.

essays L1
parliament deb.
summ, L1
acad. theses
law texts
ADV-ADV-types

23
Summary

- Highly significant register dependent ADV use in German L1 texts
- ADV-underuse in L2 data is dependent on register
- ADV-ADV-categories register dependent but do not correlate with underuse
- The generally higher frequencies in L1 essays are more difficult for learners than the lower frequencies in L1 summaries
Outlook

- Behavior of certain individual lexemes and lexeme groups (in progress; 'xxx einmal')
- More granularity than STTS offers
- More registers (ideally also spoken data)
- Underuse / overuse beyond surface statistics (syntactic categories, phrase structures)
Thank you!

- Falko is freely available at
  http://korpling.german.hu-berlin.de/falko/index.jsp
References (1/2)

References (2/2)

Examples from learner data (Falko)

1. **und [immer noch] kann man eine**
   and still can one an
   **unzufriedenheit spüren**
   dissatisfaction feel

2. **muss man [eigentlich] [nur bis ungefähr**
   must one actually only till about
   **achtzehn] überleben**
   eighteen survive

3. **Es ist [doch] [auch] statistisch belegt**
   it is also statistically proven
Error annotation and register

- Some learner data has obvious errors:
  
  *Je viel liest, desto mehr weißt* *(usb013_2006_10)*
  The much read, the more know

- Error analysis hard to apply to register:
  
  *Es kommen auch Leute nach Skandinavien nur um dort "vom Staat" zu leben. Das tolle "Staats-model" hat sich herumgesprochen, und jetzt haben die Skandinavier ein Problem.* *(hu012_2006_09)*
  People come to Scandinavia too, just to “live off the state”. Word of the cool “state-model” has gotten out, and now the Scandinavians have a problem.